# Efficient Information Retrieval through Comparison of Dimensionality Reduction Techniques with Clustering Approach

Poonam P. Rajurkar, Aditya G. Bhor, Komal K. Rahane, Neha S. Pathak
Student, Department of Information Technology

Anagha N. Chaudhari
Asst. Professor, Department of Information Technology
Pimpri Chinchwad College of Engineering,
Pune, India.

## ABSTRACT
In day today life huge amount of electronic data is generated from various resources. Such data is literally large and not easy to work with for storage and retrieval. This type of data can be treated with various efficient techniques for cleaning, compression and sorting of data. Preprocessing can be used to remove basic English stop-words from data making it compact and easy for further processing; later dimensionality reduction techniques make data more efficient and specific. This data later can be clustered for better information retrieval. This paper elaborates the various dimensionality reduction and clustering techniques applied on sample dataset C50test of 2500 documents giving promising results, their comparison and better approach for relevant information retrieval.

## Keywords
High Dimensional Datasets, Dimensionality reduction, SVD, PCA, Clustering, K-means.

## 1. INTRODUCTION
For the complex data sets there is a problem in retrieval of the necessary information from particular records. As the original datasets are multidimensional in nature, this dimensionality is a threat for handling of data. The data usually needs retrieving of specific information as per user needs, so for retrieving the specific information from this really high dimensional datasets, it needs reduction in dimensionality. Hence, for this there are different dimensionality reduction techniques and by using these techniques the datasets initially are reduced to lower dimensional space, and then processed further. In later stages the dataset is clustered using k-means clustering approach , to group similar data together for better retrieval and storage.

Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are used for dimensionality reduction and further obtained outputs from both are applied with the K-means clustering.

**Modules**
Module 1: Preprocessing

Module 2: Application of Dimensionality Reduction Techniques
Module 3: Applying Clustering Approaches

## 2. SYSTEM ARCHITECTURE

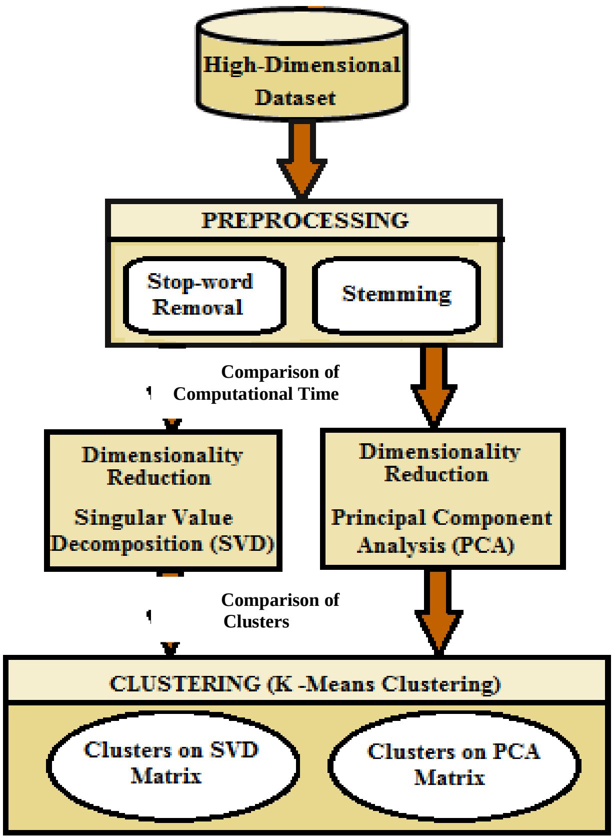


**Figure 1 : System Architecture**

## 2.1 Module 1: Pre-processing
Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects [2]. If there is much irrelevant and duplicate information present or noisy and unreliable data, then knowledge discovery gets more difficult [2].

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. As there are many stop words in a given text file at any given instance, these words increase the dataset size and also slows the further processing of data mining techniques [1].The data preprocessing techniques used in this paper are stop word removal and stemming.". The purpose of both this method is to remove various suffixes, to reduce number of words, to have exactly matching stems, to save memory space and time [5].

## 2.2 Module 2: Application of Dimensionality Reduction Techniques

DR techniques are proposed as a data preprocessing step. This process identifies a suitable low dimensional representation of previous data [3]. Dimensionality Reduction (DR) in the dataset improves the computational efficiency and accuracy in the analysis of data. The problem of dimension reduction can be defined mathematically as follows : given a *r*-dimensional random vector **Y**=(*y*1,*y*2,….*yr*)T, it's main objective is to find a representation of lower dimension **P**=(*p*1,*p*2,…,*pk*)T, where *k*<*r*, which preserves the content of the original data, according to some criteria [4].

Dimensionality reduction is the process of reducing the number of random variables under some consideration. A word matrix (documents*terms) is given as input to reduction techniques like Principal Component Analysis(PCA) and Singular Value Decomposition(SVD).

Dimensionality Reduction is done when:

- Irrelevant features exist in data.
- High dimensional data visualization.

## 2.3 Singular Value Decomposition(SVD)

In data mining, this algorithm can be used to better understand a database by showing the number of important dimensions and also to simplify it, by reducing of the number of attributes that are used in a data mining process [10]. This reduction removes unnecessary data that are linearly dependent in the point of view of Linear Algebra [10]. In computational science, it is commonly applied in Information Retrieval (IR).SVD can be implemented using formula shown below.

$$\mathbf{A}_{[m \times n]} = \mathbf{U}_{[m \times k]} * \sum_{[k \times k]} * (\mathbf{V}_{[k \times n]})^{T}$$

where,

**A**: *m x n* matrix (m documents, n terms)

**U**: *m x k* matrix (m documents, k concepts)

**Σ**: *k x k* diagonal matrix (strength of each 'concept')

**V**: k

## 3. PRINCIPAL COMPONENT ANALYSIS(PCA)

In principal component analysis we find the directions in the data with the most variation , i.e. the eigenvectors corresponding to the largest eigen values of the covariance matrix, and project the data onto these directions [9].PCA is an analysis tool for identifying patterns in data and expressing these data in such a way that it highlights their similarities and differences.

PCA is unsupervised algorithm. PCA ignores the class labels in the datasets.This algorithm is used to find the direction that maximizes the variance in the datasets.

Algorithm:

1. Organise data into n*m matrix where m is measurement type and n is number of samples.
2. Subtract off mean from each measurement type.
3. Calculate Covariance Matrix.
4. Calculate Eigen Values and Eigen Vectors from the Covariance Matrix

## 3.1 Module 3 :Applying Clustering Approaches

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters) [6]. Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. The most popular clustering technique is k-means clustering. K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships [7]. The k-means is one of the simplest clustering techniques and it is commonly used in data mining, biometrics and related fields [8].

Euclidean Distance Formula for K-means implementation is-

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\| x_i - v_j \|)^2$$

where,

'$|x_i - v_j|$' is the Euclidean distance between $x_i$ and $v_j$.

'$c_i$' is the number of data points

in $i^{th}$ cluster.

'$c$' is the number of cluster centers.

'$x_i$ is the data points in $i^{th}$ cluster.

'$v_j$' is the center of $j^{th}$ cluster.

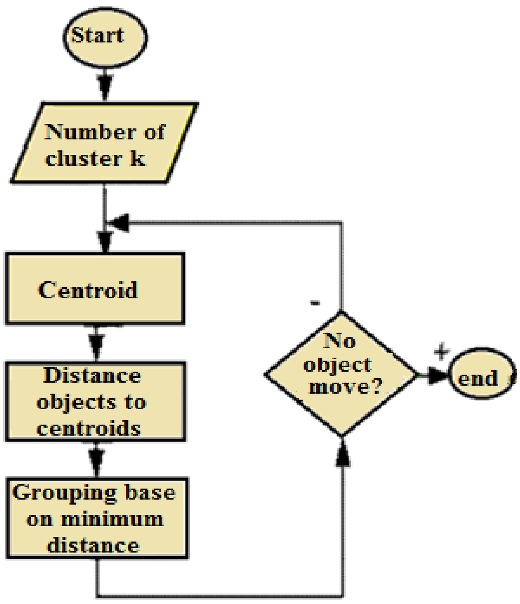


**Figure 2 : K-means Flowchart**

K-means algorithm

Let X = {x1,x2,x3,……..,xn} be the set of data points and

V = {v1,v2,…….,vc} be the set of centers.

1. Randomly select *'c'* cluster centres [11].
2. Calculate the distance between each data point and cluster centre [11].
3. Assign the data point to the cluster centre whose distance from the cluster centre is minimum of all the cluster centres [11].
4. Recalculate the new cluster centre using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, *'$c_i$'* represents the number of data points in $i^{th}$ cluster [11].

5. Recalculate the distance between each data point and new obtained cluster centres [11].
6. If no data point was reassigned then stop, otherwise repeat from step 3 [11].

## 4. EXPERIMENTAL RESULTS

C50test dataset was used for performing all the experiments [12]. It contains 2500 files which were preprocessed using dimensionality reduction techniques like SVD and PCA. Dimensionally reduced word matrix was then clustered using Clustering technique like K-means.

### 4.1 SVD Matrix With Computation Time

The dataset obtained after preprocessing is treated with SVD algorithm to produce a dimensionality reduced word matrix as shown in Figure 3, the time required for this computation was 1.8513 seconds.

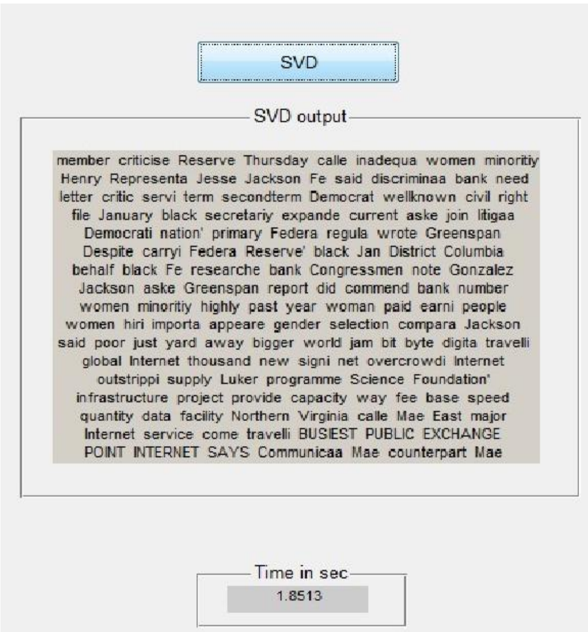


**Figure 3 : SVD Computation**

### 4.2 PCA Matrix with Computation time

The dataset obtained after preprocessing is treated with PCA algorithm to produce a dimensionality reduced word matrix as shown in Figure 4, the time required for this computation was 1.1524 seconds.

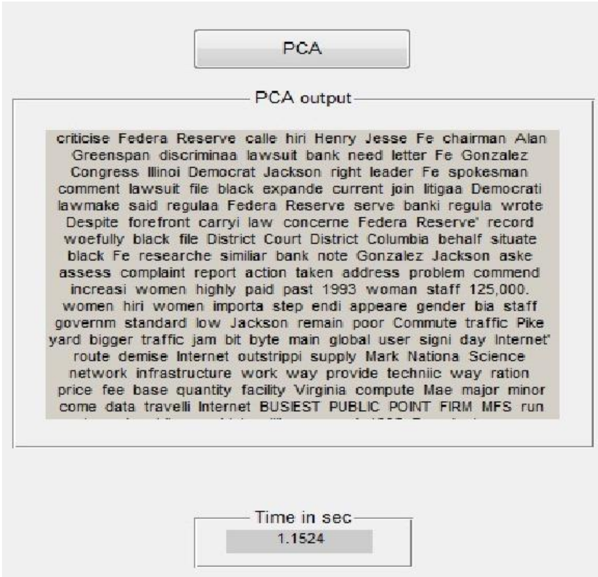


**Figure 4 : PCA Computation**

### 4.3 Retreival Of Word From SVD AndPCA Matrices

The dimensionally reduced word matrices obtained by SVD and PCA were used for retreiving words from it through an user interface by entering user desired query (Figure 5). The retrieval time for matching user query and displaying its path in dataset documents was less for PCA, proving it better than SVD.


**Figure 5 : Query retrieval**

### 4.4 Clusters formed using k-means Technique on SVD results

The output in the form of dimensionally reduced word matrix by applying SVD to the preprocessed dataset was clustered using K-means clustering, by considering k=4, i.e. 4 clusters were formed(Figure 6).

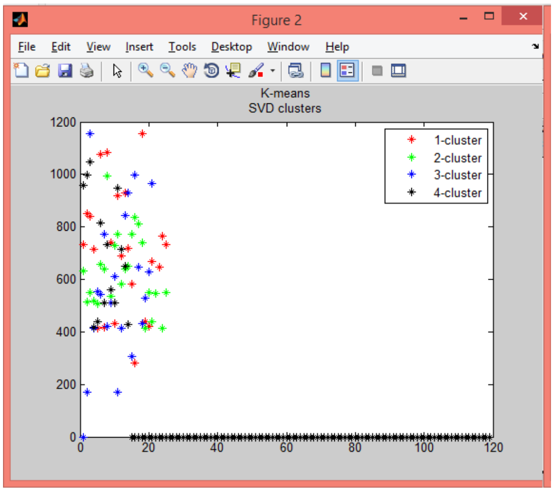


**Figure 6 : SVD Data Clusters**

### 4.5 Clusters Formed Using K-Means Technique On PCA Results

Dimensionally reduced word matrix obtained by applying PCA on preprocessed dataset was clustered using K-means clustering forming 4 clusters which were prominently categorized(Figure 7).

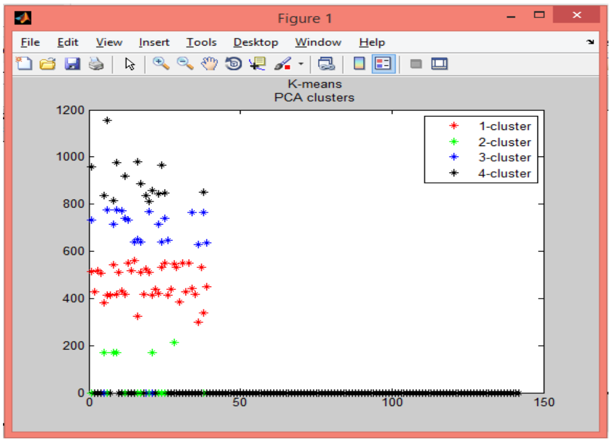


**Figure 7 : PCA Data Clusters**

## 5. EXPERIMENTAL ANALYSIS (SVD V/s PCA)

Both the techniques were applied to pre-processed C50test dataset. The observations made depict that PCA being the next version of SVD proves to be better in many ways, i.e. the computational time of PCA technique for formation of dimensionally reduced weight matrix was less than that of SVD as well as the retrieval time for retrieving any specific word in the dimensionally reduced word matrix through user query was significantly less for PCA algorithm(Figure 8).

And further when computations of PCA and SVD were clustered using K-means Clustering algorithm, the clusters formed on SVD data were scattered and not prominent where as those formed on PCA data were accurate, categorized and better than SVD.

**Table 1.Comparison of SVD & PCA based on various Experimental Parameters**

| Parameters | SVD | PCA |
|---|---|---|
| Computational time(sec) | 1.851 3 | 1.1524 |
| R1etrieval time(sec) | 0.452 9 | 0.3527 |
| Reduced Matrix size (doc x max_terms ) | 2500 x 289 | 2500 x 290 |
| Clusters Formed | Scatte red | Categorize d and Accurate |

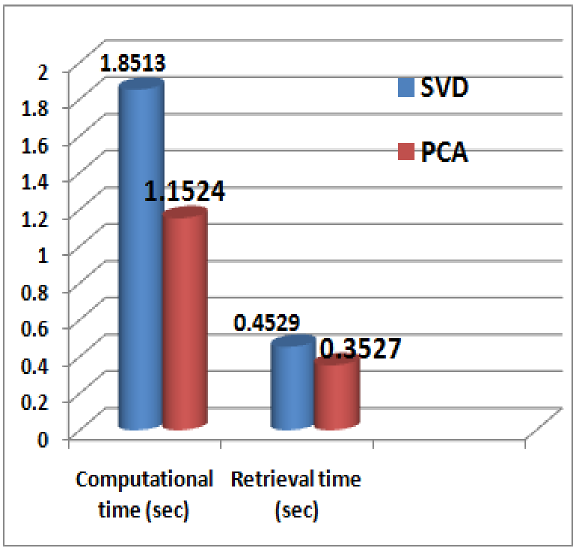


**Figure 8 : SVD and PCA comparison graph**

## 6. CONCLUSION

This paper mainly contributes to provide a new approach with comparative study of the dimensionality reduction techniques as SVD and PCA to improve the performance of clustering. Users can make comprehensive choices among various available dimensionality reduction techniques referencing this study. The main objective is to achieve best performance of K-means clustering by treating original dataset like C50test with pre-processing techniques like stop-word removal followed by stemming and then later with dimensionality reduction techniques. Hence after obtaining SVD-PCA computations as SVD-K-means, PCA-K-means; it proves that K-means done on data computed by PCA technique gives accurate, categorized and better results than SVD. This research has shown comparable results of available techniques on C50test dataset and a better approach for relevant Information Retrieval system. Future work would consist of several experiments to be performed with different dimensions and datasets achieving better accuracy and developing an Information Retrieval system for accurate and relevant document retrieval in less possible time.

## 7. REFERENCES

[1] V. Srividhya, R. Anitha , " Evaluating Preprocessing Techniques in Text Categorization ",ISSN 0974-0767,International Journal of Computer Science and Application Issue 2010.

[2] Nguyen Hung Son ,"Data Cleaning and Data Preprocessing".

[3] Lei Yu Binghamton University, Jieping Ye ,Huan Liu ,Arizona State University, "Dimensionality Reduction for data mining-Techniques, Applications and Trends".

[4] Ch. Aswani Kumar ,"Analysis of Unsupervised Dimensionality Reduction Techniques" , ComSIS Vol. 6, No. 2, December 2009.

[5] C.Ramasubramanian, R.Ramya, "Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm" , International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013.

[6] Rui Tang, Simon Fong, Xin-She Yang, Suash Deb," Integrating nature-inspired optimization algorithms to k-means clustering", 978-1-4673-2430-4/12/$31.00 ©2012 IEEE.

[7] Carlos Cobos, Henry Muñoz-Collazos, RicharUrbano-Muñoz, Martha Mendoza, Elizabeth Leónc, Enrique Herrera-Viedma "Clustering Of Web Search Results Based On The Cuckoo Search Algorithm And Balanced Bayesian Information Criterion " ELSEVIER Publication, 2014 Elsevier Inc. All rights reserved ,21 May 2014

[8] Agnihotri, D.; Verma, K.; Tripathi, P., "Pattern and Cluster Mining on Text Data," Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on, vol., no., pp.428,432, 7-9 April 2014.

[9] Rasmus Elsborg Madsen, Lars Kai Hansen and Ole Winther ,"Singular Value Decomposition and Principal Component Analysis" , February 2004.

[10] https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Dimensionality_Reduction/Singular_Value_Decomposition

[11] https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

[12] http://archive.ics.uci.edu/ml/datasets/Reuter_50_50